

DOCUMENT RESUME

ED 300 450

TM 012 499

AUTHOR Green, Donald Ross
TITLE Local versus National Calibrations.
PUB DATE Apr 87
NOTE 15p.; Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 20-24, 1987).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Achievement Tests; Grade 4; Grade 6; Intermediate Grades; Latent Trait Theory; National Norms; Program Evaluation; Reading Comprehension; *Reading Tests; School Districts; *Standardized Tests; *Test Construction; Testing Problems; Testing Programs; Test Reliability; Test Validity
IDENTIFIERS *Calibration; Comprehensive Tests of Basic Skills; Curriculum Based Assessment; Philadelphia City Wide Testing Program; *Test Customization

ABSTRACT

An assessment of the Philadelphia (Pennsylvania) City-Wide Testing Program is presented to determine the effects of using customized achievement tests to achieve multiple purposes. The tests are derived from, or related to, the tests in a nationally standardized test battery, but include items written specifically to measure the local curriculum. It is generally believed that item response theory can resolve the problems resulting from this test construction process. The reading and mathematics tests developed for Philadelphia contain items calibrated on the reading and mathematics scales of the Comprehensive Tests of Basic Skills (CTBS), Form U. Data for grades 4 and 6 for 1984-86 on the reading comprehension test were chosen for the study. Results of the analysis indicate that some fairly small effects of the local curricula were found. Thus, the effect of calibrating the Philadelphia items and adding them to the CTBS set accomplished its intended purpose--to increase reliability without any serious consequences for validity. Five data tables conclude the document. (TJH)

* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *

ED300450

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)

☒ This document has been reproduced as
received from the person or organization
originating it.

☐ Minor changes have been made to improve
reproduction quality.

• Points of view or opinions stated in this docu-
ment do not necessarily represent official
OERI position or policy.

"PERMISSION TO REPRODUCE THIS
MATERIAL HAS BEEN GRANTED BY

DONALD ROSS GREEN

TO THE EDUCATIONAL RESOURCES
INFORMATION CENTER (ERIC)."

Local versus National Calibrations

by

Donald Ross Green

CTB/McGraw-Hill

Paper presented at the 1987 Annual Meeting of the American
Educational Research Association, Washington D.C.

TM012499

All or almost all achievement test programs in school systems serve as a source of information for several contending parties: students, parents, teachers, principals, curriculum supervisors, superintendents, and often state and/or federal agencies as well. Thus either the tests used must serve several purposes and provide several kinds of data or several different tests must be given.

Generally speaking tests are most likely be valid when built for a specific purpose and used only for that purpose; this principle suggests using several different tests. However, as everyone knows, that option meets with strong resistance from all concerned (except perhaps test publishers).

In recent years people have begun to try to resolve this conflict by creating what have come to be called customized tests. Customized tests are typically derived from, or related to, the tests in a nationally standardized test battery but include some (or many) items written specifically to measure the local curriculum. The basic idea is to serve many purposes in a single test. Thus, on the one hand, the intent is to have a test that measures just what has been taught (teachers want this) or what was supposed to be taught (supervisors and principals want this), while on the other hand the goal is to have national norms for the scores (the public, parents, newspapers, school boards, and school administrators usually want this).

There are some problems that arise in attempting to achieve each of these goals in a single test but it is widely believed that item response theory solves these problems. The reason for this belief is that under certain circumstances item response theory does indeed accomplish this goal; however the nature of these circumstances is not always understood. Thus the limitations imposed on the inferences that can be drawn from the scores of tests that match what has been taught are often ignored (Mehrens & Green, 1986); similarly, the use of a set of calibrated items to estimate norms instead of the use of regular normed instrument may also have some logical problems (Yen, Green, & Burket, 1987).

One of the problems of this latter sort can arise when some locally calibrated items are added to (or used instead of) nationally calibrated items. If the local calibration has occurred before or near the beginning of use of a new local curriculum, and if, as is likely, the locally calibrated items were written to fit the local curriculum, instruction with the new local curriculum is more likely to affect the parameters of the locally calibrated items than those of the nationally calibrated items.

Since the conditions just described fit the situation that has existed in Philadelphia, that program provides a chance to look at some data to see just how strong these effects may be. It is the purpose of this study to determine if the potential problem just described has arisen in the Philadelphia City-Wide Testing Program.

The Philadelphia City-Wide Tests have been constructed to provide information for a variety of purposes, some of which require normative information and some of which require criterion-referenced information. In order to keep the length of the tests short enough to be administered in one class period it seemed desirable to have as many items as possible that could provide both kinds of data. Thus many of the tests consist of items coming from two quite different sources. In both reading and mathematics at least twenty of the items at any grade level were drawn from the CTB/McGraw-Hill pool of nationally calibrated items (hereafter the CTB items); therefore scores from these sets of items can be used to obtain norms estimates. The remaining items in the tests were written specifically by and/or for Philadelphia to measure the objectives of the newly installed system curricula (hereafter the Philadelphia items).

However in the interests of reliability and better measurement at more points on the scale (a twenty item test has many limitations) it was judged desirable to have items do double duty wherever possible. Therefore using the data from the Fall 1984 administration of the tests, those Philadelphia items that appeared to fit the scale were calibrated and used in getting the norms estimates.

In the Fall of 1985 and again in the Fall of 1986 these same sets of items were used again. Because at these points the students had been instructed for one or two years with the new curriculum installed in the fall of 1984, it becomes possible that

the Philadelphia items will yield different scores in these latter years than the CTB items which were not written to fit the Philadelphia curriculum. This is the result that would be expected if the new Philadelphia curriculum represented a departure from their prior curriculum and if it was indeed implemented in the classrooms.

Procedure

The reading and mathematics tests developed for Philadelphia contain items calibrated on the reading and mathematics scales of the Comprehensive Tests of Basic Skills, Form U published in 1981. The Philadelphia tests were examined to determine how many of them had adequate numbers of both CTB and Philadelphia calibrated items to make comparisons of scores reasonable. Two tests were found: the grade 4 and the grade 6 Reading Comprehension tests. The remaining tests had less than 10 Philadelphia calibrated items; ten was judged to be the minimum number of items necessary to provide an adequately valid estimate of group performance on the trait.

Since the numbers of students taking these test was very large, a sample created by taking every tenth case was used. Descriptive statistics for these tests are shown in Table 1.

Table 2 shows these same statistics separately for the two components, the CTB items and Philadelphia items and also the correlations between the two sets of scores. Note that both scores are estimates of student status on the trait but since

they were not divided by content these estimates are fully equivalent only to the degree that the test is truly unidimensional; if achievement was truly unidimensional across all stages of learning none of the problems being examined here would arise and the year-to-year changes in the differences would have been merely error of measurement. Whether or not that is the case is the central issue.

Since the correlations are reasonably high relatively small differences between the two mean scores might be considered significant. However it does not seem appropriate to compare the means of the scores from these components. This is because large numbers of scores fall at the floors of these abbreviated tests and because the scale scores assigned to chance level (or perfect) performances, albeit rational, are somewhat arbitrary (cf. Yen, 1983). For example, the frequency distribution for the two sets of scores for grade 4 in 1984 are shown in Table 3*. It can be seen that almost ten percent of the scores fall at the floor. Therefore the comparisons made are between quartile points especially the medians.

*In 1984 the tests were given two weeks later than in 1985 and 1986. Therefore all the 1984 scale scores have been adjusted to make them comparable to the others.

Results

Table 4 shows the median and quartile values and their standard errors for the two items sets for each test each year. The standard errors are derived from the standard deviation using a multiplier which assumes a normal distribution. As was apparent from the means, scores from all these sets of items have been increasing from year to year. Their corresponding national percentiles, also shown in Table 4, give an indication of the amount of this growth.

The questions of greatest interest concern whether the Philadelphia items yield higher or lower estimates than the CTB items and whether this relationship changes over the years. Table 5 shows these differences at the several quartiles. If the instruction during these two years with the new standard curriculum has differentially affected performance on the two sets of items one might expect that the difference between the medians for the Philadelphia and CTB items would become larger each year. Except at the 25th percentile in grade 4 the 1984 to 1986 changes are all in this direction. The two strongest trends occur at the first quartile but in opposite directions for the two grades. In grade six all the changes from year to year except at the 75th percentile for 1985 to 1986 are in this direction. However at grade 4 the progression is not regular for any of the quartiles.

Discussion

Clearly the data indicate a less than crystal clear trend in the year to year changes in the differences in performances between the two sets of scores. In grade six there is a reasonably solid pattern of year to year growth in the difference but the amounts at Q1 and Q2 are not large.

Any number of reasons for the lack of a strong trend can be offered. It may be that Philadelphia curriculum in reading has not really changed and/or that the changes that have occurred are not strongly reflected in the content represented by the items that were calibrated. It should be noted that not all the items written by or for Philadelphia were calibrated. It may be that the new elements of the curriculum have yet to be well implemented and that in another year or two clearer trends would emerge; after all we all know that change usually occurs slowly in large school systems. Still another possibility is that the new and the old Philadelphia curricula are equivalent subsets of the national curriculum represented by CTBS, Form U.

In sum it appears that some effects of the local curricula were found but where they occurred these effects were fairly small. Thus the effect of calibrating the Philadelphia items and adding them to the CTB set accomplished what was intended, i.e., increased reliability without any serious consequences for validity. Whether or not this conclusion would continue to hold in the future is unknown. In any case it follows that growth shown is real and is not an artifact of test customization.

References

Mehrens W. A. & Green, D.R. (1986). Standardized tests and school curricula. Unpublished manuscript.

Yen, W.M., Green, D.R., & Burket, G.B. (1987). Valid normative data from customized achievement tests. Educational Measurement: Issues and Practice, 6, 7-13.

Yen, W.M., (1983). Use of the three-parameter logistic model in the development of a standardized achievement test. In R.K. Hambleton (Ed.) Applications of Item Response Theory. Vancouver, British Columbia: Educational Research Institute of British Columbia.

TABLE 1
SAMPLE MEANS AND MEDIANS

<u>YEAR</u>	<u>NO. ITEMS</u>	<u>N</u>	<u>MEAN</u>	<u>S.D.</u>	<u>S.E. MEAN</u>	<u>MEDIAN</u>	<u>S. MED</u>
GRADE 4							
1984	47	1949	611.1	106.5	2.4	628.3	3.0
1985	47	1089	619.2	101.7	3.1	635.0	3.9
1986	46	1399	636.3	91.3	2.4	649.2	3.1
GRADE 6							
1984	45	1960	686.5	85.0	1.9	695.5	2.4
1985	45	1093	698.3	80.4	2.4	706.7	3.0
1986	44	1167	704.3	76.4	2.2	711.8	2.8

TABLE 2
MEANS, MEDIANS, AND CORRELATIONS FOR THE CTB AND
PHILADELPHIA SUBSETS OF ITEMS

<u>YEAR</u>	<u>ITEM SET</u>	<u>NO. ITEMS</u>	<u>MEAN</u>	<u>S.D.</u>	<u>MEAN</u>	<u>MEDIAN</u>	<u>S.E. MEDIAN</u>	<u>r</u>
GRADE 4								
1984	CTB	30	609.9	106.2	2.4	629.4	3.015	0.76
	PHIL	17	619.8	112.5	2.5	628.9	3.194	
1985	CTB	30	616.2	106.4	3.2	636.0	4.041	0.72
	PHIL	17	621.0	106.2	3.2	630.3	4.033	
1986	CTB	30	629.8	97.3	2.6	647.7	3.260	0.73
	PHIL	17	640.2	102.2	2.7	649.1	3.424	
GRADE 6								
1984	CTB	22	687.2	79.3	1.8	699.8	2.245	0.78
	PHIL	23	686.0	96.2	2.2	698.6	2.723	
1985	CTB	22	689.3	70.7	2.1	701.6	2.680	0.78
	PHIL	23	693.1	89.6	2.7	702.5	3.397	
1986	CTB	22	701.8	74.3	2.2	710.1	2.726	0.80
	PHIL	22	705.7	86.0	2.5	711.4	3.155	

TABLE 3
FREQUENCY DISTRIBUTION
1984 GRADE 4

<u>CTB ITEM GROUP</u>				<u>PHILADELPHIA ITEM GROUP</u>			
<u>SCALE SCORE</u>	<u>FREQ</u>	<u>CUM FREQ</u>	<u>CUM PERC</u>	<u>SCALE SCORE</u>	<u>FREQ</u>	<u>CUM FREQ</u>	<u>CUM PERC</u>
395	188	188	9.6	405	179	179	9.2
426	57	245	12.6	436	79	258	13.2
493	71	316	16.2	511	119	377	19.3
523	81	397	20.4	548	125	502	25.8
543	73	470	24.1	575	133	635	32.6
559	78	548	28.1	598	151	786	40.3
572	90	638	32.7	620	192	978	50.2
584	60	698	35.8	643	181	1159	59.5
596	71	769	39.5	666	167	1326	68.0
606	58	827	42.4	692	194	1520	78.0
616	61	888	45.6	722	160	1680	86.2
625	71	959	49.2	758	139	1819	93.3
634	59	1018	52.2	803	85	1904	97.7
643	75	1093	56.1	855	32	1936	99.3
652	80	1173	60.2	890	13	1949	100.0
661	80	1253	64.3				
670	94	1347	69.1				
679	94	1441	73.9				
689	81	1522	78.1				
700	90	1612	82.7				
713	90	1702	87.3				
728	88	1790	91.8				
749	80	1870	95.9				
786	50	1920	98.5				
820	29	1949	100.0				

TABLE 4
QUARTILES FOR THE CTB AND PHILADELPHIA ITEM SETS

<u>YEAR</u>	<u>ITEM SET</u>	<u>Q1</u>	<u>NP Q1</u>	<u>MEDIAN</u>	<u>NP MEDIAN</u>	<u>Q3</u>	<u>NP Q3</u>	<u>SE Q</u>	<u>SE MEDIA</u>
GRADE 4									
1984	CTB	552.4	15	629.4	35.7	684.8	64.4	3.25	3.01
	PHIL	556.3	16	628.9	35.5	695.6	70.3	3.44	3.19
1985	CTB	560.1	17	636.0	38.5	689.1	67.1	4.35	4.04
	PHIL	561.5	17	630.3	36.2	694.4	69.7	4.34	4.03
1986	CTB	583.7	22	647.7	44.4	694.2	69.6	3.51	3.26
	PHIL	583.4	21	649.1	45.1	709.1	76.7	3.69	3.42
GRADE 6									
1984	CTB	648.0	16	699.8	31.4	742.2	54.2	2.42	2.24
	PHIL	642.1	15	698.6	30.9	742.6	54.6	2.93	2.72
1985	CTB	656.6	18	701.6	32.2	737.3	51.2	2.89	2.68
	PHIL	656.1	18	702.5	32.5	746.9	57.5	3.66	3.40
1986	CTB	663.2	20	710.1	36.1	750.9	59.9	2.94	2.73
	PHIL	665.7	20	711.4	36.7	755.1	62.6	3.40	3.16

TABLE 5
DIFFERENCES IN SCALE SCORES FOR CTB
AND PHILADELPHIA ITEM SUBSETS

	Q1	MEDIAN	Q3
GRADE 4			
1984	3.9	-0.5	10.8
1985	1.4	-5.7	5.3
1986	-0.3	1.4	14.9
GRADE 6			
1984	-5.9	-1.2	0.4
1985	-0.5	0.9	9.6
1986	2.5	1.3	4.2